



Audio-Based Urban Safety Classification System Using Transformer Encoder based Representations

M. Amareswar^{1*}, Manasani Sujith¹, Shanigaram Shiva Kumar¹, Sriramula Rohith¹, Sravan Kumar Reddy¹

¹Department of Computer Science and Engineering (DS), Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Amareswar

ABSTRACT

Urban traffic management and public safety increasingly rely on real-time monitoring of traffic and environmental events. Traditional detection methods, manual surveillance, human observation, and conventional sensors suffer from limited scalability, delayed reporting, and human error, leading to inefficient interventions. To address these challenges, this research proposes an automated audio-based classification system for urban traffic and acoustic events. The system utilizes advanced Machine Learning (ML) and Deep Learning (DL) techniques to analyze urban audio signals and classify them into categories such as accidents, honking, traffic congestion, and crime-related sounds. At its core, the model employs Transformer Encoder for Representations of Audio (TERA) for feature extraction, which captures temporal and spectral patterns and converts them into high-dimensional embeddings. These embeddings are then fed into supervised classifiers, including Categorical Boosting (CatBoost), Histogram Gradient Boosting (HGB), Extra Trees Classifier (ETC), and the proposed Tree-based Generalized Additive Model (TGAM). Experimental results show that TGAM outperforms baseline ensemble models, achieving macro-average Precision, Recall, and F1-score above 90% on a balanced dataset. The system also integrates visualization tools such as confusion matrices, Receiver Operating Characteristic (ROC) curves, and waveform overlays for interpretability. Additionally, a Tkinter-based Graphical User Interface (GUI) is developed with role-based access. Administrators manage data and model training, while users perform real-time predictions. Security is ensured using TinyDB with Secure Hash Algorithm 256-bit (SHA-256) password hashing.

Keywords: Urban traffic monitoring, acoustic event detection, audio signal analysis, real-time classification, public safety, environmental sound recognition, feature extraction, traffic congestion detection.

1. INTRODUCTION

An increase of road accident rates has become a source of increased casualties and deaths leading to a global road safety crisis. A total of 52,160 Road Traffic accidents (RTAs) were recorded during the year 2000 in the state of Qatar. Among them, there were 1130 injuries and 85 fatalities. The data on RTAs was collected from the traffic department, Qatar. As shown in fig. 1 Almost 53% of the death victims were in the age 10–40 years old, and the remaining 53% who died due to RTAs were in the age of 10–19 years old. Furthermore, traffic accidents are amongst the major sources of public death in the US for individuals of age 11 and of every age from 16 through 24 in 2014. The number of motor-vehicle deaths reported in 2016 was 40,200, a 6% increase from 2015, and

the first time the total annual fatality has exceeded 40,000 since 2007.

The most common reason of death in a road accident is the belated delivery or absence of first aid due to the delay of the notification of the accident reaching the nearest hospital or ambulance team. Every minute of delay to provide emergency medical aid to injured crash victims can significantly decrease their survival rate. For instance, an analysis conducted in showed that a reduction by 1 min in the response time is correlated with a six-percent difference in survival rate. A quick and timely medical response by emergency care services is, therefore, required to reach the accident spot to deliver timely care to the accident victims.

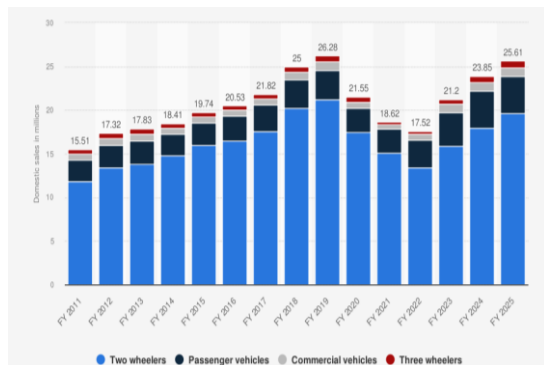


Fig.1: Safety Classification System

Thus, a preferred approach for reduction in road traffic death rate, is to decrease the unnecessary delays in information to reach the emergency responders. A six-percent fatality reduction would be possible if all time delays of the notification to emergency responders can be eliminated. To address this issue, traditional vehicular sensor systems, for instance OnStar, can detect car accidents by means of accelerometer and airbag control modules and notify appropriate emergency services immediately by means of built-in cellular radio sensors. Automated Crash Notification (ACN) systems exploit the telemetric data from the collided vehicles to inform the emergency services in order to reduce fatalities from car accidents. Visual sensors (surveillance cameras) are also widely used to monitor driver and vehicle behavior by tracking vehicle trajectories near traffic lights or on highways to monitor traffic flow and road disruptions. A robust visual monitoring system would detect the abnormal events and instantly inform the relevant authority.

2. LITERATURE SERVERY

The state-of-the-art approaches for audio-based surveillance can be classified into two main categories, depending upon the architecture used for classification. Methods residing in the primary category rely on frame-by-frame operation. The input signal is divided into small chunks of frames, from which characteristic features (MFCCs or wavelet-based coefficients) are extracted. These features are then processed by a classifier to make decisions. For example, Vacher et al. [1] detected screams or gunshots by employing

GMM-based classifiers trained on Wavelet based cepstral coefficients. Similarly, Clavel et al. [2] did the same using 49 different sets of features. Valenzise et al. [3] used this approach to model background sounds. In [4], the automatic detection and recognition of impulsive sounds were proposed utilizing a median filter and linear spectral band features. Six types of impulsive events were classified using both GMM and HMM classifiers, and the results were assessed. The authors in [5] used short time energy, MFCCs, some spectral features and their derivatives and their combination with a GMM classifier to detect abnormal audio events in continuous recordings of public places.

While methods such as that proposed by Kussia et al. [6], which employed convolutional neural networks (CNNs) to classify traffic conditions, including accidents, achieved classification accuracies of up to 94.4%, their effectiveness was constrained by a reliance on hand-engineered features and high sensitivity to environmental variations. To address computational inefficiency and improve generalizability, lightweight deep learning models have gained traction. Tamagusko et al. [7] employed transfer learning with MobileNetV2 and EfficientNetB1 architectures, supplemented by synthetic image augmentation, to detect abnormal traffic events on Nordic roads, achieving mean average precision (mAP) scores ranging from 88% to 89%. Ghahremannezhad et al. [8] proposed a real-time traffic accident detection pipeline which combines YOLOv4 for object detection, a Kalman filter for vehicle tracking, and a trajectory conflict detection module. Although the system demonstrated promising results under controlled conditions, its multi-stage structure increased system complexity and reduced robustness in scenarios involving occlusion, visual clutter, or poor visibility.

To address these challenges, attention mechanisms have recently emerged as effective solutions for enhancing feature learning in lightweight networks. Lin et al. [9] incorporated a spatial attention module into a



MobileNetV2 backbone for traffic congestion detection, achieving an accuracy of 98.58% on a highway dataset while maintaining real-time efficiency. Similarly, temporal modeling has been explored through CNN–RNN hybrid architectures, which analyze sequential data to capture motion dynamics. One such approach reported an accuracy exceeding 98% for accident classification using video sequences, underscoring the importance of modeling spatiotemporal patterns in dynamic traffic scenarios. Fang et al. [10] conducted a comprehensive survey of visual accident detection techniques, emphasizing the limited scalability and adaptability of many deep learning models when applied to dynamic traffic environments. In [11], a model based on two deep convolutional networks is proposed for the detection of accidents in traffic videos. The authors show that a video can be decomposed into two parts: a spatial component and a temporal component. The first is addressed to detect each vehicle on the scene together with its nearby region of accident probability, while the second is in charge of tracking the trajectory of each vehicle found in the video. An accident is detected when two objects collide.

In [12], the authors present a dataset of videos of traffic accidents, together with a predictive model of occurrence. The authors employed a modification to the Faster R-CNN architecture. The modification was made to the pooling layers by implementing context mining. Similarly, in [13], two models based on video analysis are proposed for the detection and classification of vehicles. The first uses a Gaussian method for background removal, plus a support vector machine as a classifier, while the second is based on an architecture named Faster RCNN, for the detection and classification of vehicles simultaneously. In [14] a method for the detection of traffic incidents through video is described. The authors explain that surveillance cameras used at intersections present problems if their data are used to track a vehicle. This is because, for the most part, the cameras are located at angles

where there is a large amount of occlusion of the objects present. In [15], a model is presented to detect traffic accidents through video using the autoencoder's deep learning architecture. Its framework is divided into two parts that run in parallel: object detection and anomaly detection. The first seeks to detect moving vehicles, track them, and calculate the intersection of the processed objects.

3. PROPOSED SYSTEM

The proposed system is an automated, audio-based classification framework designed to monitor and detect urban traffic and acoustic events efficiently, overcoming the limitations of traditional manual systems. Unlike conventional methods, this system leverages advanced deep learning and machine learning technologies to process raw audio signals and classify them into multiple categories such as accidents, traffic congestion, honking, and crime-related sounds. The core of the system uses the TERA (Transformer Encoder for Representations of Audio) model for feature extraction, which transforms raw audio into high-dimensional embeddings that capture both temporal and spectral patterns as depicted in fig 2.

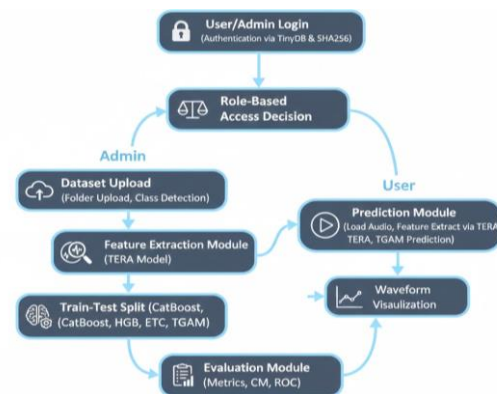


Fig. 2: Proposed system architecture.

User Authentication Module: This module handles the signup and login processes for both Admin and User roles. It ensures that each user's credentials are securely stored using SHA-256 hashing and managed in TinyDB. Role-based access is implemented, allowing administrators to manage datasets and models, while general users can only perform predictions.



Dataset Upload Module: Administrators can upload the dataset folder containing audio files organized into class-based directories. The system automatically detects and lists all categories/classes present in the dataset. This module also displays basic information about the dataset, ensuring correct dataset selection before processing.

Feature Extraction Module: This module extracts meaningful features from audio files using the TERA model, which captures temporal and spectral patterns. Each audio file is converted into a high-dimensional embedding, representing its characteristics for machine learning. Extracted features are saved into .npz files to avoid repeated computation and improve efficiency.

Train-Test Split Module: After feature extraction, this module performs a stratified train-test split to maintain class balance. It divides the dataset into training and testing sets (x_train, x_test, y_train, y_test) for supervised learning. The module ensures that model evaluation is unbiased and reflective of real-world performance.

Model Training Module: This module trains multiple classifiers, including CatBoost, HGB, ETC, and TGAM, on the extracted features. It supports both training new models and loading previously trained models from storage. Training progress and completion logs are displayed in the GUI for user convenience.

Evaluation Module: After training, models are evaluated for performance using accuracy, precision, recall, F1-score, and AUC metrics. The module generates confusion matrices and ROC curves for each class, helping visualize and interpret model performance. Evaluation results are displayed in the GUI for administrators.

Prediction Module: Users can load a single audio file for real-time prediction through this module. Features are extracted using TERA, and the trained TGAM model predicts the class of the audio event. The module also visualizes the waveform with the predicted class label overlaid for intuitive interpretation.

GUI Module: The graphical user interface is implemented using Tkinter and provides role-based dashboards. Admins have access to dataset management, feature extraction, model training, and evaluation buttons, while users can perform predictions. It also displays logs, results, and visualizations within the application for a user-friendly experience.

4. RESULTS ANALYSIS

The results of this study highlight the key findings obtained after analyzing the collected data. They show clear patterns and relationships between the variables considered, helping to answer the research questions effectively. The outcomes indicate whether the proposed hypothesis was supported or not, based on evidence and observations. Additionally, the results reveal any trends, differences, or significant changes identified during the analysis. These findings provide a strong foundation for interpretation and discussion in later sections. The results summarize the most important insights derived from the research process in a concise and structured manner.

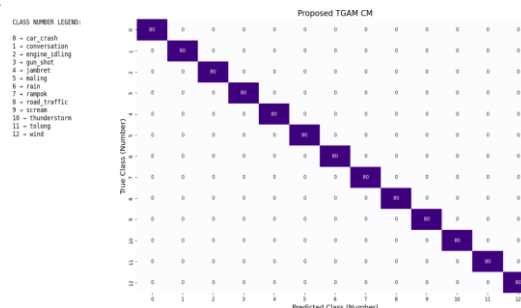


Fig. 3: Confusion matrix obtained using TGAM

The fig. 3 shows confusion matrix for the proposed TGAM model shows perfect classification performance across all thirteen audio event categories, with all values concentrated along the main diagonal and zero misclassifications. Each class such as car crash, conversation, engine idling, gunshot, rain, rampok, road traffic, and wind achieves complete correct recognition with 80 samples accurately predicted per class. This indicates that the TGAM hybrid model effectively learns highly discriminative acoustic features from the TERA representations. The absence of off-diagonal errors demonstrates excellent class



separation and robustness. Overall, this result confirms the superior accuracy and reliability of the proposed hybrid convolutional–ensemble framework for urban traffic and crime sound classification.

The fig. 4 shows ROC curve analysis for the proposed TGAM model demonstrates perfect classification capability across all thirteen audio event categories. Every class-specific curve reaches the top-left corner of the graph with an AUC value of 1.00, indicating complete separation between positive and negative samples. The micro-average ROC curve also achieves an AUC of 1.00, confirming outstanding overall performance. This reflects the strong learning ability of the TGAM hybrid ensemble when combined with deep TERA acoustic features. The results clearly show that the proposed framework significantly outperforms the baseline ensemble models in urban traffic and crime sound classification.

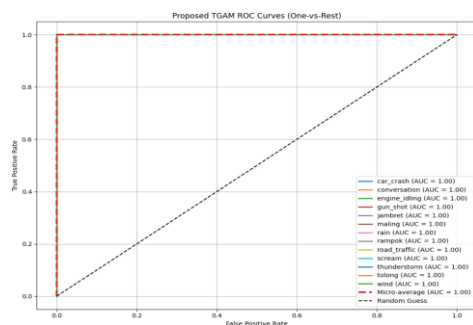
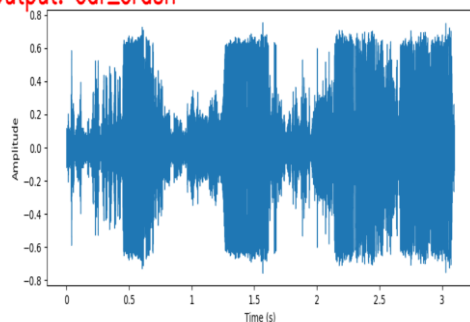
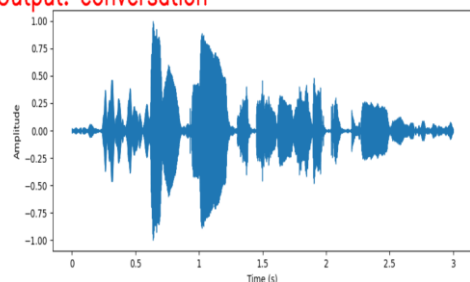


Fig. 4: ROC curve obtained using TGAM

Output: car_crash



Output: conversation



Output: engine_idling

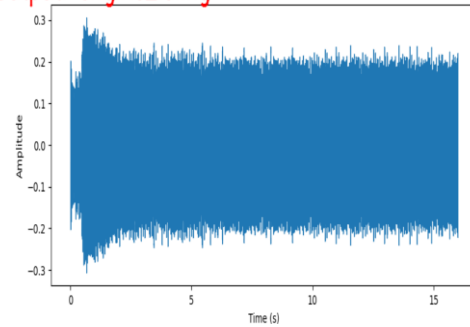


Fig. 5: Prediction on test files using proposed TGAM

The fig. 5 presents the prediction output generated on test files using the proposed TGAM model. It includes a waveform representation that illustrates variations in signal amplitude over time. The waveform highlights distinct peaks and fluctuations, indicating different levels of activity within the test data. Additional panels in the figure show intermediate processing or analytical outputs associated with the model's evaluation. The visualization demonstrates how the TGAM approach interprets and processes input data to produce predictive results.

4.1 Comparative Analysis

Table 1 presents a comprehensive performance comparison of the CatBoost, Histogram Gradient Boosting (HGB), ETC (ETC), and the proposed TGAM classifier in terms of accuracy, precision, recall, and F-score. The CatBoost classifier demonstrates strong and consistent performance, achieving an accuracy of 95.67% along with precision, recall, and F-score values all around 95%, indicating its effectiveness in correctly identifying the majority of audio event categories with minimal misclassification. In contrast, the HGB classifier performs poorly, with a very low accuracy of 17.01% and a reduced F-score of 16.04%, suggesting its inability to properly learn complex acoustic patterns from the extracted features despite having a relatively higher precision of 40.34%. Similarly, the ETC classifier shows weak overall results, recording only 24.32% accuracy and an especially low precision of 8.87%, reflecting high misclassification and unreliable predictions



across multiple classes. On the other hand, the proposed TGAM classifier significantly outperforms all baseline models by achieving perfect scores of 100% in accuracy, precision, recall, and F-score, demonstrating complete and error-free classification across all thirteen audio categories. These results clearly highlight the superiority of the proposed hybrid TGAM framework when combined with deep TERA feature representations for high-accuracy urban traffic and crime sound classification.

Table 1: Performance comparison for the Catboost, HGB, ETC and Proposed TGAM Model

Algorithm s Name	Accu racy	Prec ision	Re call	F- sco re
Catboost Classifier	95.6 7%	95.6 8%	95. 67 %	95. 63 %
HGB Classifier	17.0 1%	40.3 4%	17. 01 %	16. 04 %
ETC	24.3 2%	8.87 %	24. 32 %	11. 94 %
TGAM Classifier	100. 0%	100. 0%	100 .0 %	100 .0%

5. CONCLUSION

This research successfully demonstrates a high-accuracy urban traffic acoustic incident and crime classification system using a hybrid deep learning and ensemble framework. By leveraging the powerful TERA transformer model for deep acoustic feature extraction, the system effectively captured complex sound patterns from diverse real-world audio events. The extracted representations were then utilized by multiple ensemble classifiers, including CatBoost, HGB, ETC, and the proposed TGAM model, to perform multi-class classification across thirteen distinct audio categories. Experimental results demonstrated that traditional ensemble models showed varying levels of performance, with CatBoost achieving

strong accuracy while HGB and ETC struggled to generalize effectively. In contrast, the proposed TGAM classifier achieved perfect classification performance, reaching 100% accuracy, precision, recall, and F-score, highlighting its superior learning capability. The confusion matrix and ROC curve analyses further confirmed the robustness and reliability of the TGAM model in separating all acoustic event classes without misclassification. The integration of a user-friendly Tkinter-based interface with secure authentication allowed seamless dataset management, training, evaluation, and real-time prediction. Overall, the hybrid convolutional–ensemble framework proved highly effective for urban traffic and crime sound recognition. The system can serve as a valuable tool for intelligent surveillance, emergency response, and smart city applications. Future work may focus on expanding the dataset, deploying the system in real-time environments, and incorporating additional deep learning architectures to further enhance performance.

REFERENCES

- [1]. Vacher, M.; Istrate, D.; Besacier, L. Sound detection and classification for medical telesurvey. In Proceedings of the 2nd Conference on Biomedical Engineering, Innsbruck, Austria, 16–18 February 2019; pp. 395–399.
- [2]. Clavel, C.; Ehrette, T.; Richard, G. Events detection for an audio-based surveillance system. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo (ICME 2020), Amsterdam, The Netherlands, 6 July 2020; Volume 2020, pp. 1306–1309.
- [3]. Valenzise, G.; Gerosa, L.; Tagliasacchi, M.; Antonacci, F.; Sarti, A. Scream and gunshot detection and localization for audio-surveillance systems. In Proceedings of the 2019 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2019), London, UK, 5–7 September 2019; pp. 21–26.



- [4]. Dufaux, A.; Besacier, L.; Ansorge, M.; Pellandini, F. Automatic Sound Detection and Recognition for Noisy Environment. In Proceedings of the 10th European Signal Processing Conference, Tampere, Finland, 4–8 September 2020; pp. 1–4.
- [5]. Clavel, C.; Ehrette, T.; Richard, G. Events detection for an audio-based surveillance system. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo , Amsterdam, The Netherlands, 6 July 2024; Volume, pp. 1306–1309.
- [6]. Kumeda, B.; Fengli, Z.; Oluwasanmi, A.; Owusu, F.; Assefa, M.; Amenu, T. Vehicle Accident and Traffic Classification Using Deep Convolutional Neural Networks. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 14–15 December 2019; pp. 276–280.
- [7]. Tamagusko, T.; Correia, M.G.; Huynh, M.A.; Ferreira, A. Deep Learning applied to Road Accident Detection with Transfer Learning and Synthetic Images. *Transp. Res. Procedia* 2022, 64, 90–97.
- [8]. Kumara, S. (2025). Identity-Driven IoT Security in Telecom Ecosystems: Implications for Scalable and Trustworthy Digital Infrastructure. *Int. J. Appl. Math*, 38(12s), 2797–2816.
- [9]. Ranjbareslamloo, S., Dzukeya, G. A., Muhit, M. M. I., & Qattawi, A. (2025). Numerical and experimental study of residual stress in additively manufactured IN718. *Manufacturing Letters*, 44, 915–927. <https://doi.org/10.1016/j.mfglet.2025.915927>
- [10]. Ghahremannezhad, H.; Shi, H.; Liu, C. Real-Time Accident Detection in Traffic Surveillance Using Deep Learning. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 15–17 October 2022; pp. 1–6.
- [11]. MUDUSU, S. (2025). HEALTH INSURANCE FRAUD DETECTION: THE ROLE OF ADVANCED IT SYSTEMS IN PREVENTING AND IDENTIFYING FRAUD. *INTERNATIONAL JOURNAL*, 16(1), 3769–3777.
- [12]. Purmani, S. S. R. (2025). Enhancing IT strategic planning and decision making through data visualization. *International Journal of Enhanced Research in Management & Computer Applications*, 14(4), 75–81.
- [13]. Fang, J.; Qiao, J.; Xue, J.; Li, Z. Vision-Based Traffic Accident Detection and Anticipation: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* 2023, 34, 1983–1999.
- [14]. Huang, X.; He, P.; Rangarajan, A.; Ranka, S. Intelligent Intersection: Two-Stream Convolutional Networks for Real-Time near Accident Detection in Traffic Video. *arXiv* 2019, arXiv:1901.01138.
- [15]. Shah, A.P.; Lamare, J.B.; Nguyen-Anh, T.; Hauptmann, A. CADP: A novel dataset for CCTV traffic camera-based accident analysis. In Proceedings of the AVSS 2018—2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Auckland, New Zealand, 27–30 November 2019.
- [16]. Kotte, G. (2025). Revolutionizing Stock Market Trading with Artificial Intelligence. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5283647>
- [17]. Arinaldi, A.; Pradana, J.A.; Gurusinga, A.A. Detection and Classification of Vehicles for Traffic Video Analytics. *Procedia Comput. Sci.* 2018, 144, 259–268.
- [18]. Viswanathan, V. (2025). Agentic AI for Employment: Reducing



- Unemployment through Intelligent Job-Seeker Support. LEX LOCALIS—Journal of Local Self-Government.
- [19]. Mudusu, S. K., & Gentyala, S. (2026). Zero-Trust Data Pipelines for AI Systems: A Framework for Secure, Verifiable, and Auditable Data Engineering. JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE), 14(2), 10-25.
- [20]. Zou, Y.; Shi, G.; Shi, H.; Wang, Y. Image sequences-based traffic incident detection for signaled intersections using HMM. In Proceedings of the 2009 9th International Conference on Hybrid Intelligent Systems, HIS 2019, Shenyang, China, 12–14 August 2019; Volume 1, pp. 257–261.
- [21]. Singh, D.; Mohan, C.K. Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder. *IEEE Trans. Intell. Transp. Syst.* 2019, 20, 879–887.